University of Zurich[UZH]

http://r-bayesian-networks.org/
gilles.kratzer@math.uzh.ch
sonja.hartnack@access.uzh.ch

GILLES KRATZER, APPLIED STATISTICS GROUP, UZH

SONJA HARTNACK, VETSUISSE, UZH

ECVPH WORKSHOP, ZURICH 7-9 MAY 2019

RISK FACTOR ANALYSIS

OUTLINE

- Risk factor analysis introduction
- p-value model selection / change of estimate
- Theory on model selection/variable selection/feature exctraction
    - Machine learning
    - step AIC
    - Random Forest
        - Ensemble method
        - Decision tree
        - Variable importance
    - varrank
        - Relevance/redundancy
        - Mutual infromation/entropy

RISK FACTOR ANALYSIS

- **RFA** used for guiding diagnosis, therapy or disease control
- A **risk factor** is any **attribute**, characteristic or exposure of an individual that **change** the **likelihood** of developing a disease/exposure or injury/condition
  - Classical example in epidemiology: age, gender, underweight/obesity, unsafe sex, high blood pressure, tobacco, alcohol consumption, and unsafe water, sanitation and hygiene, breed, managment, housing system …

- **RFA** could be **litterature based**
- **RFA** could be **data driven** (model predictive based)
  - This process is usually considered as a problem of variable selection
  - Controversial!
  - No unique strategy
  - No clear strategy

SPIRIT OF AUTOMATED VARIABLE SELECTION

- **Risk factors** are variables that **influence** the outcome **significantly**
- **Risk factor** are **important** for modelling
- **Risk factors** are not **confounders**
- **Within modelling:** risk factors = covariates

- **model prediction** is about:
  - **causal links** requires **interventions/experiments**
  - **observed** associations

- From **observational** data:
  - associations only!
  - … still underlying causal links
  - what is **important**? **effect size**?
  - what is **significant**? at **individual** level? at **population** level?
  - risk of **subgroup** vanishing effect
  - model validation?
  - supervised/**unsupervised**
  - **training/testing** datasets

NAIVE APPROACHE

- **Important** covariates = **significant** p-values!
  - No because test hypothesis
  - **Unaccounted** multiple testing
  - Complex dependencies among each other
  - Testing order? Search algorithm?
- **Change of estimate**
  - Model building strategy?
  - What is a large change? Scaling?

NAIVE APPROACHE

```
glm(formula = casecontrol ~ age + gender + eatbeef + eatpork +
    eatveal + eatlamb + eatpoul + eatcold + eatveg + eatfruit +
    eateggs + slt_a + dlr_a + dlr_b, family = binomial(link = logit),
    data = salm)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10757  -0.50183  -0.17426  -0.00019   1.94506

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.233e+01  3.956e+03   0.003   0.9975
age          6.627e-03  2.592e-02   0.256   0.7982
gender1      1.514e-01  8.690e-01   0.174   0.8617
eatbeef1    -9.155e-01  9.235e-01  -0.991   0.3216
eatpork1     1.169e+00  1.426e+00   0.820   0.4122
eatveal1     3.863e+00  1.722e+00   2.244   0.0248 *
eatlamb1    -1.200e+01  2.780e+03  -0.004   0.9966
eatpoul1     2.632e+00  1.192e+00   2.208   0.0272 *
eatcold1    -1.525e+01  3.956e+03  -0.004   0.9969
eatveg1     -2.596e+00  4.332e+00  -0.599   0.5490
eatfruit1   -2.489e+00  1.210e+00  -2.057   0.0397 *
eateggs1     2.319e+00  1.320e+00   1.756   0.0791 .
slt_a1       3.642e+00  1.442e+00   2.526   0.0115 *
dlr_a1       2.321e-01  1.029e+00   0.226   0.8215
dlr_b1      -4.901e-01  1.692e+00  -0.290   0.7721
```

- **Vocabulary**: Variable selection = feature extraction = predictor selection
  - *Task: Selecting one model from a set of possible models*
- **Machin learning (ML):**

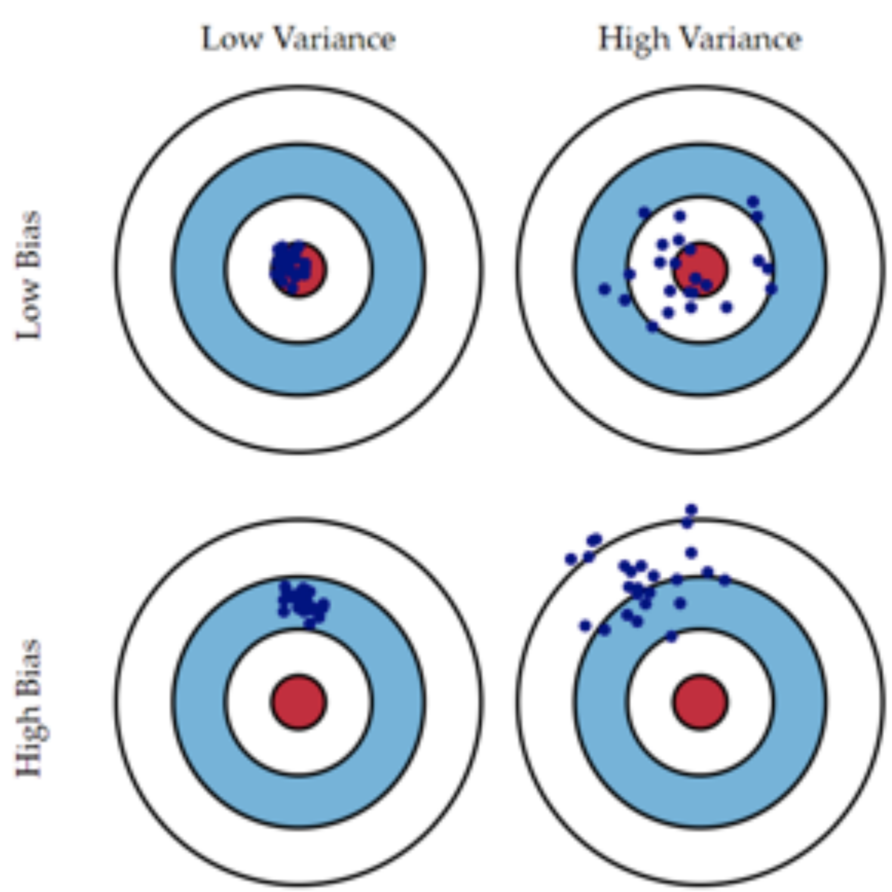
MODEL SELECTION: STEPAIC

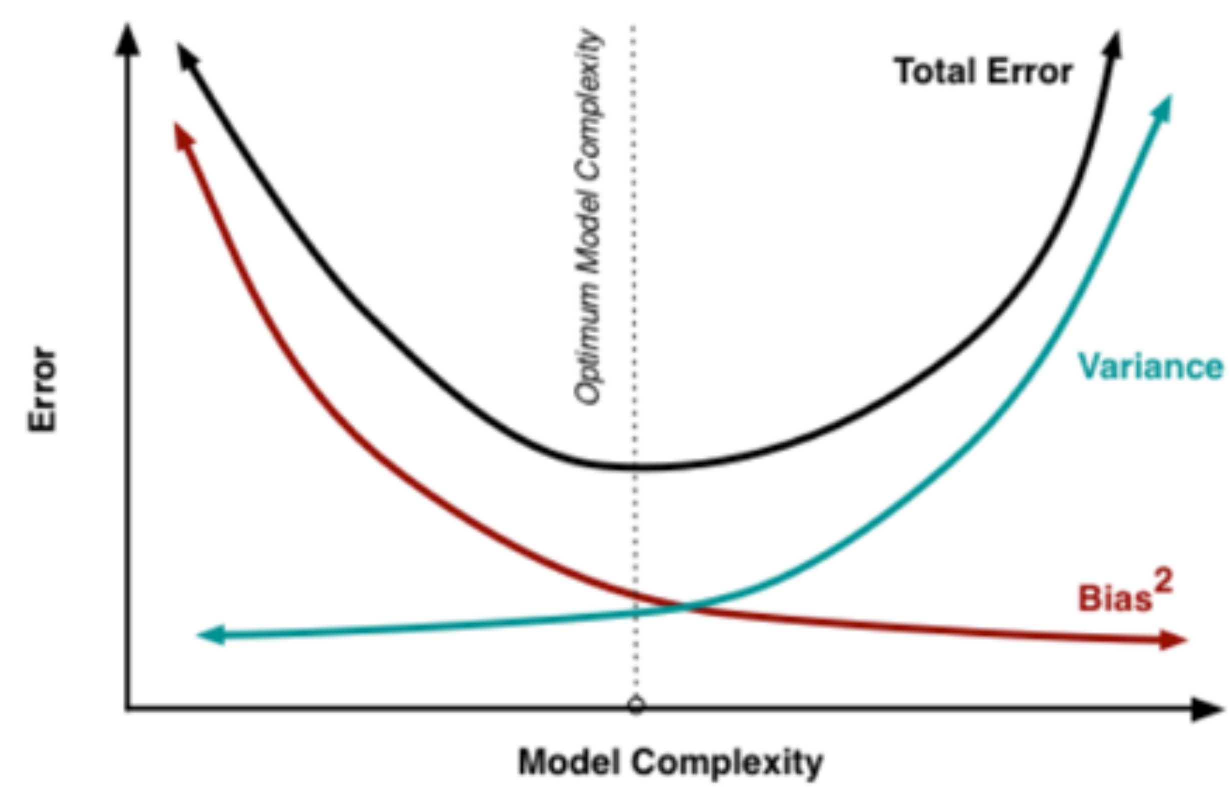
- **StepAIC**:
- The concept of model complexity can be used to create measures aiding in model selection
- Scores that deal with this trade-off between goodness of fit and model simplicity
    - Akaike information criterion (AIC)

$$AIC = 2k - 2\hat{L}$$

    - Bayesian information criterion(BIC)

$$BIC = ln(n)k - 2\hat{L}$$





MODEL SELECTION: STEPAIC

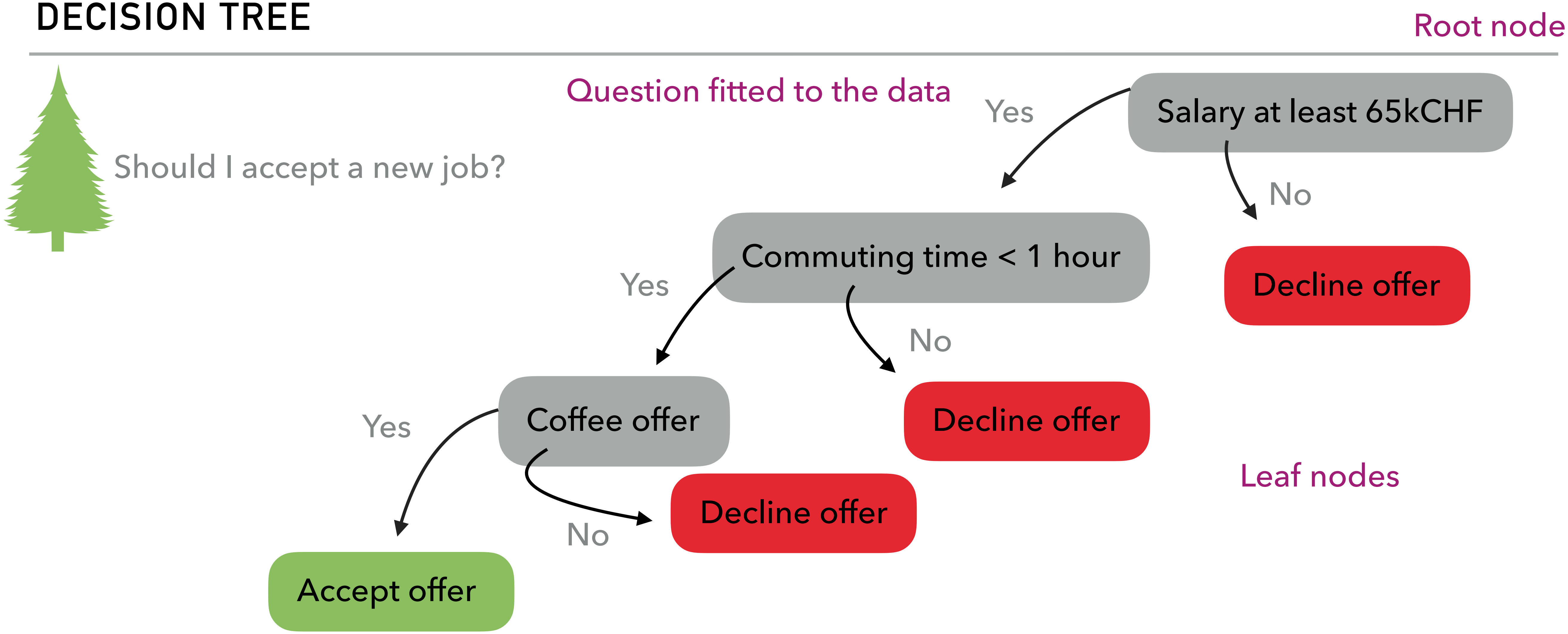
```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
casecontrol ~ age + gender + eatbeef + eatpork + eatveal + eatlamb +
    eatpoul + eatcold + eatveg + eatfruit + eateggs + slt_a +
    dlr_a + dlr_b

Final Model:
casecontrol ~ eatbeef + eatveal + eatpoul + eatfruit + eateggs +
    slt_a


       Step Df   Deviance Resid. Df Resid. Dev      AIC
1                              58   44.10115 74.10115
2 - eatlamb  1 0.01527492      59   44.11643 72.11643
3  - gender  1 0.03080419      60   44.14723 70.14723
4   - dlr_a  1 0.04430816      61   44.19154 68.19154
5     - age  1 0.03502305      62   44.22656 66.22656
6 - eatcold  1 0.13259298      63   44.35915 64.35915
7   - dlr_b  1 0.13632402      64   44.49548 62.49548
8 - eatpork  1 0.61677047      65   45.11225 61.11225
9  - eatveg  1 1.28606341      66   46.39831 60.39831
```

RANDOM FOREST

- **Random forests** or random decision forests are an **ensemble** learning method for classification, regression, variable selection
- Operates by constructing a **multitude** of **decision trees** at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

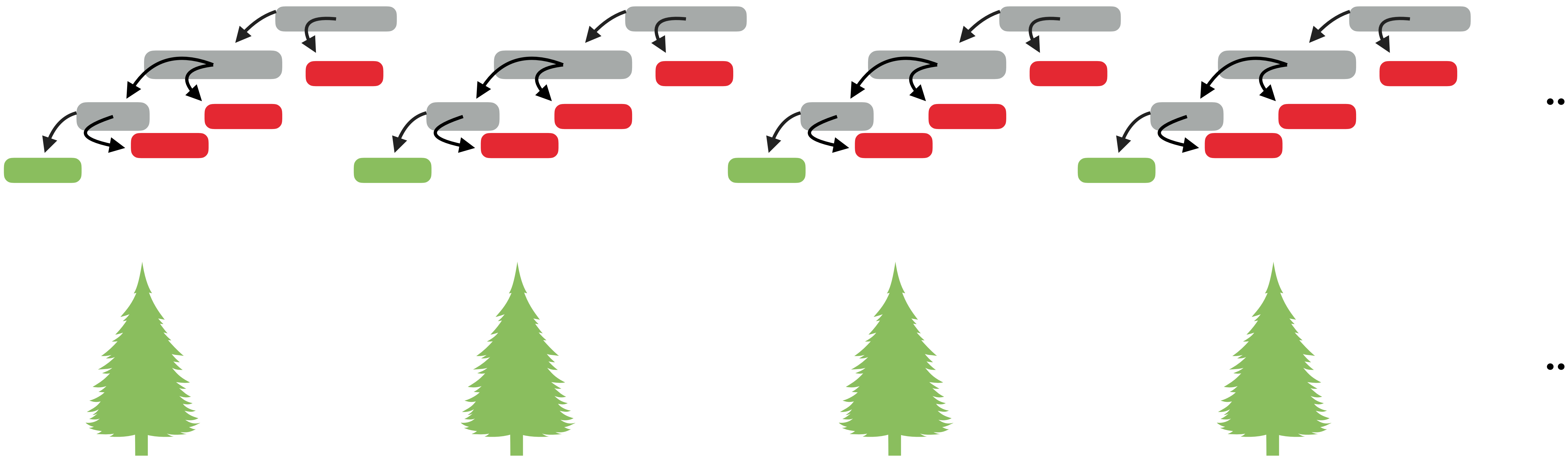
- **Ensemble methods**:
  - Ensemble methods are **meta-algorithms** that combine several machine learning techniques into one model in order to decrease variance (bagging), bias (boosting) or improve predictions (stacking)
  - **bagging** = bootstrap aggregation: Reduce the variance of an estimate in averaging multiple estimates (later)
  - **boosting**: combining weak model (slightly better than random guess) into strong model

DECISION TREE



- **Root node**
  - Entry point to a collection of data
- **Inner nodes**
  - A question (statistical dependency) is fitted to the data
- **Leaf nodes**
  - Correspond to the decision to take (or conclusion to make) if reached
- **Pruning**
  - To avoid over-fitting of learning data
  - To achieve a trade-of between prediction accuracy and complexity

RANDOM FOREST AND VARIABLE IMPORTANCE

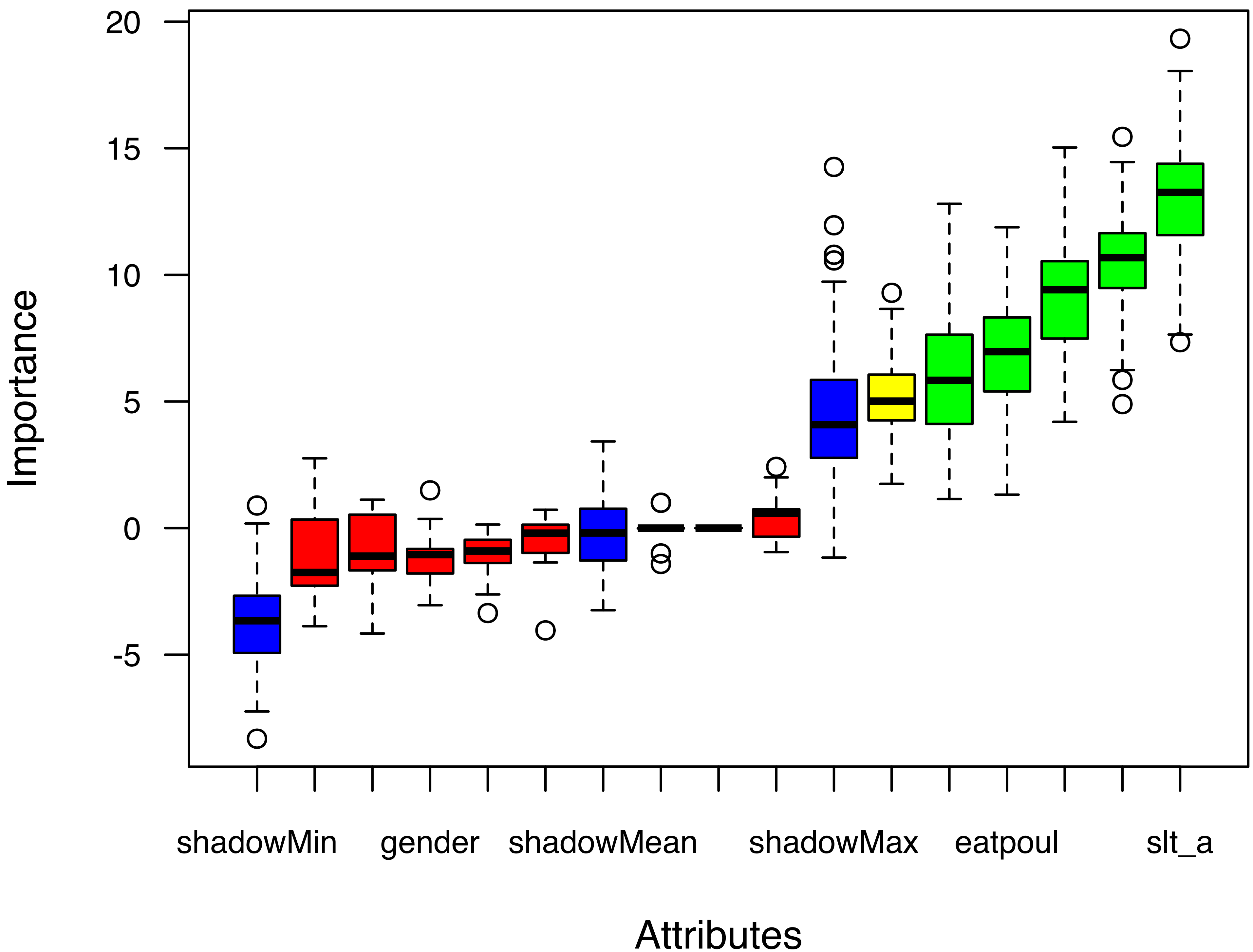


- From a **single tree** to **random forest**:
  - Training data is sampled from the full data set with replacement
  - Subset of variables is considered when deciding how to split each node
  - Fitted/traied until the leaf nodes contain one or very few samples

- **Disadvantages** of random forests
  - Random forests improvment on single decision trees but more sophisticated techniques: gradient-boosted trees
  - A forest is less **interpretable** than a single decision tree
  - Generating forest may require significant memory usage for storing trees

- **Advantages** of random forests
  - No tuning parameters
  - Tend not to overfit the data
  - Exctract general patterns within the data and reduce sensitivity to noise
  - Ability to handle non-linear numeric and categorical predictors and outcomes
  - Predictor variable importance can be computed

RANDOM FOREST AND VARIABLE IMPORTANCE



RANDOM FOREST AND VARIABLE IMPORTANCE

- **Boruta:**
- The dataset is extended by adding copies of all variables (remove any corellation with the response variable)
- Random forest classifier is run on the whole data set and Z-scores are computed for all attributes (another importance measure)
- Out of all shadow attributes find the one with the maximum Z score and then assign a hit to every attribute that scored better
- For each attribute with undetermined importance perform a two-sided test of equality with the the one obtained for shadow attribute with maximum Z-score
- Mark the attributes which have importance signifcantly lower than the shadow with maximum Z-score as `unimportant' and permanently remove them from the data set
- Remove all shadow, artificially added attributes repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

University of Zurich[UZH]

System epidemiology

- Typically the set of possible variables is formidable
  - The classical approach for variable selection is based on prior scientific knowledge (29%)[1]
  - Change of estimate (18%)[1]
  - Stepwise model selection (16%)[1]
- No prior model
- Not one outcome experiment

varrank Variable ranking for better time allocation

- Variable ranking based on a set of variable of importance
- Model free. Based on information theory metrics
- Mixture of variables (continuous and discrete). Discretisation through rule/clustering
- Ranking of 100 variables with 100'000 observations in ~14 minutes! (forward greedy search)

[1] *Walter et al (2009)*

varrank

Score = Relevance - Redundancy / Normalization

Outcome

Highly relevant variable

Redundant group of variable

Other covariate

f_i candidate feature to be ranked

C set of variables of importance

S set of already selected variables

$$\mathrm{H}(X) = \sum_{n=1}^{N} P(x_n) \log P(x_n)$$

$$\mathrm{MI}(X;Y) = \sum_{n=1}^{N} \sum_{m=1}^{M} P(x_n; y_m) \log \frac{P(x_n; y_m)}{P(x_n) P(y_m)}$$

$$\mathrm{score}_i = \mathrm{MI}(f_i; \mathbf{C}) - \beta \sum_{F_s \in \mathbf{S}} \alpha(f_i, f_s, \mathbf{C})\ \mathrm{MI}(f_i; f_s)$$

Estévez and al. (2009)

$\beta = 1/|\mathbf{S}|$ and $\alpha(f_i, f_s, \mathbf{C}) = \dfrac{1}{\min(\mathrm{H}(f_i), \mathrm{H}(f_s))}$

MAXIMUM RELEVANCE MINIMUM REDUNDANCY

f_i candidate feature to be ranked

C set of variables of importance

S set of already selected variables

$$\mathrm{H}(X) = \sum_{n=1}^{N} P(x_n) \log P(x_n)$$

Average amount of information of one RV

$$\mathrm{MI}(X;Y) = \sum_{n=1}^{N} \sum_{m=1}^{M} P(x_n; y_m) \log \frac{P(x_n; y_m)}{P(x_n)P(y_m)}$$

Mutual dependence between two RV

Greedy search

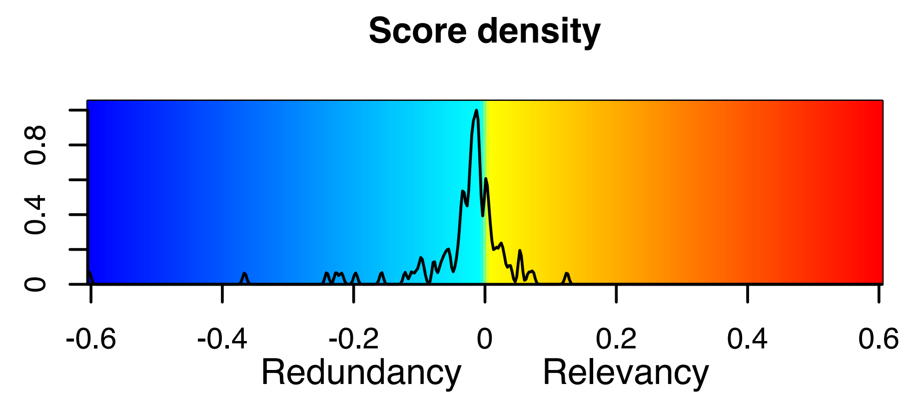
Forward - argmax

$$\mathrm{score}_i = \underbrace{\mathrm{MI}(f_i; \mathbf{C})}_{\text{Relevance}} - \beta \sum_{F_s \in \mathbf{S}} \underbrace{\alpha(f_i, f_s, \mathbf{C})}_{\text{Normalization}} \underbrace{\mathrm{MI}(f_i; f_s)}_{\text{Redundancy}}$$

Estévez and al. (2009)

$$\beta = 1/|\mathbf{S}| \text{ and } \alpha(f_i, f_s, \mathbf{C}) = \frac{1}{\min(\mathrm{H}(f_i), \mathrm{H}(f_s))}$$

MAXIMUM RELEVANCE MINIMUM REDUNDANCY



	slt_a	eatveal	eatveg	eatfruit	eatpork	eatlamb	eatcold	eatbeef	dlr_a	eateggs	gender	dlr_b	eatpoul	age
slt_a	0.125													
eatveal	0.073	0.053												
eatveg	0.03	0.025	0.024											
eatfruit	0.067	0.036	0.006	0.054										
eatpork	0.04	0.022	0.018	0.017	0.01									
eatlamb	0.011	-0.078	-0.012	-0.006	-0.004	0								
eatcold	0.028	0.015	-0.009	-0.098	-0.122	-0.035	0.002							
eatbeef	0.008	-0.046	-0.019	-0.016	-0.013	-0.009	-0.01	-0.011						
dlr_a	0.054	-0.241	-0.158	-0.197	-0.603	-0.008	-0.218	-0.065	0.002					
eateggs	0.002	0	-0.006	-0.105	-0.008	-0.078	-0.034	-0.019	-0.012	-0.011				
gender	0.002	0	-0.012	-0.012	-0.011	-0.021	-0.017	-0.017	-0.015	-0.013	-0.015			
dlr_b	0.006	-0.057	-0.038	-0.042	-0.062	-0.227	-0.017	-0.093	-0.053	-0.024	-0.018	-0.021		
eatpoul	0	-0.366	-0.098	-0.069	-0.112	-0.06	-0.039	-0.035	-0.037	-0.033	-0.03	-0.022	-0.02	
age	0.005	-0.055	-0.029	-0.034	-0.025	-0.022	-0.019	-0.017	-0.022	-0.027	-0.025	-0.032	-0.03	

TOWARD RECOMMENDATION

- Modeling should start with defendable set of assumptions
- Based on background knowledge (that a computer program typically does not possess)
  - Previous studies in the same field of research
  - Expert knowledge
  - Common sense
- Event-per-variable! Sample size/# cases
  - if <10: penalized likelihood (ridge regression)

GLOSSARY: STATISTICS VERSUS MACHINE LEARNING

Statistics	Machine learning
Fitting/estimation/selecting	Learning
Data point	Instance
Regression	Supervised learning
density estimation/clustering	Unsupervised learning
Covariate	Feature
Response/Outcome	Label
Model	Network/graph/structure

Thank you for your attention

xkcd.com